DOCUMENT RESUME

ED 271 959 FL 015 814

AUTHOR	MacWhinney, Brian, Ed.; And Others
TITLE	Access to the CHILDES Data Base, and CHILDES Transcript Coding Conventions.
INSTITUTION	Carnegie-Mellon Univ., Pittsburgh, Pa. Dept. of Psychology.
PUB DATE	Apr 86
NOTE	12p.
AVAILABLE FROM	Carnegie-Mellon University, Department of Psychology, 3328 Longbow, Pittsburgh, PA 15235.
PUB TYPE	Reference Materials - Vocabularies/Classifications/Dictionaries (134) -- Journal Articles (080)
JOURNAL CIT	Transcript Analysis; v3 n1 p99-108 Apr 1986
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Cataloging; Child Language; *Classification; *Databases; Guidelines; Indexing; *Information Retrieval; *Search Strategies; *Telecommunications; Time Sharing
IDENTIFIERS	*Coding

ABSTRACT

Two articles provide: (1) instructions for gaining telecommunications access to CHILDES, a child language database; and (2) the conventions guiding transcript coding in the database. The first article provides access instructions for telecommunications systems, the commands a user would need to know, and brief notes on searching and other aspects of file manipulation. The transcript coding conventions are for the 1.1 version of CHILDES. Included are notes on changes from the 1.0 version and instructions on coding conventions. The conventions include: auxiliary files, headers, the main textual line, dependent tiers, special symbols used to delimit utterances, and scoping or coreferencing. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED271959

TRANSCRIPT ANALYSIS

VOLUME 3 No. 1 -- 1986

Appendix I: Access to the CHILDES Data Base

Appendix II: CHILDES Transcripts Coding Conventions

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Brian MacWhinney

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL015814

Appendix I: Access to the CHILDES Data Base

It is now possible for members of the CHILDES to access the database through GTE Telenet, as well as through a variety of other telecommunications links.

The CHILDES data base is stored on a hard disk attached to a MicroVAX II at CMU. The Arpanet name of the VAX is MALLOW.PSY.CMU.EDU. If you have an account on an Arpanet machine just type "telnet a.psy.cmu.edu". When you receive the USERNAME: prompt, type MALLOW and you will be passed on through to MALLOW. Then you will again see the USERNAME: prompt and you should type in your own username and then you will see the PASSWORD: prompt and you should type in your password.

Other users can access MALLOW through GTE Telenet. Access to MALLOW.PSY.CMU.EDU via GTE Telenet is described below.

1. Find out the number of the Telenet Access Computer nearest to you by calling one of these numbers:

   800/336-0437 (from all states except Virginia, Alaska, and Hawaii)
   800/572-0408 (from Virginia)
   703/893-6026 (from outside Continental United States)

2. Dial the number.

3. Put the phone in the coupler or switch the modem to "data" (unless your modem does this automatically) and press the RETURN key twice. TELENET will then recognize the speed of your terminal and adjust accordingly. After you type two RETURN's, TELENET should type TERMINAL=at your terminal. Respond by typing "d1" for most terminals, or just hit RETURN and "d1" will be assumed as the default.

4. After you identify your terminal type, TELENET will type an "at" sign (@) on your terminal to prompt you to issue a command. You then type: c 412 670 The letter "c" means connect and the numbers are for MICOM-A at CMU.

5. At this point you should get the message CONNECTED followed by PORT GROUP?, which is the prompt from the MICOM. You would then type MALLOW and the MICOM will connect you to MALLOW.

6. When you get the CMU-PSY-MALLOW banner, and it asks for your username, type your username (usually your last name) and <CR> For a valid account name and password, please call Brian MacWhinney at (412) 268-3793.

7. When it asks for your password, type your password and <CR> Then you are logged on to MALLOW.

8. Continue with your job on MALLOW.

9. When you are all done, you type LOGOUT in the normal way and then type RETURN@RETURN, that is carriage return, at sign, carriage return. You should then get TELENET's @ sign again, at which point you type D for disconnect. You can then hang up the phone.

Once logged on to MALLOW, you will need to learn these systems:

1. The basic commands of the operating system. The operating system is version 4.0 of VAX/VMS. The commands themselves are DCL or DEC Command Language. By typing HELP TOPICS you will see a listing of the most important commands.

2. The local additions to the operating system called "logsub." Type HELP LOGSUB to read about these.

3. The EMACS text editor which has many functions that can help in data analysis. Once within EMACS, type "escape-X learn" for a tutorial.

In order to reach the CHILDES data, type: go dbase<CR>. We will try to maintain a file called "00newusers.doc" in the [dbase] directory which includes helpful hints for new users.

Path specifications under VMS have this shape:

host::device:[user.subdirectory.subdirectory...]filename.ext;generation

as in

mallow::psy$user:[dbase.small.bloom.allison]allison1.txt;3

This would give you the third version or generation of the file allison1.txt in the "allison" subdirectory of the "bloom" subdirectory of the "small" subdirectory of the "dbase" directory on the disk device called "psy$user" on MALLOW. Usually, you will not have to mention the device name or the node name MALLOW. However, you will need to become familiar with using subdirectory names and file names. You move around using the various logsub commands. To read files you use the "type" command. To edit them you use the "emacs" command, as in:

$ emacs allison1.txt

The simplest way to do fast searches for strings and creation of concordances is to use VMS "search" command. For information on how to use the command, type "help search". You will find that the /output switch is necessary if you want the output to go to a file and that the /window switch is useful for extracting several lines before and after the key word.

You will find that you do not have permission to write out new versions of the files to the [dbase] directory. However, you can write out versions to your home directory. Within EMACS, you use the command C-X, C-W or "write-named-file" and make sure that you give the full specification for a file in your home directory, such as [holland.work]new.txt

It may be possible to "download" files back to your local micro if you have a sufficiently powerful communications program. One program that will work for this is Kermit, which we run on MALLOW.

For advanced users, we have available FORTRAN, PASCAL, FranzLISP, and BASIC. By typing "u csh" you can enter the c-shell in a simulated UNIX called Eunice where you can run and compile C programs. The HUM C-programs and various grep scripts run in the c-shell.

If you would like to examine the data over a fast direct line using more powerful facilities, we invite you to come to Pittsburgh for a visit. Call us to make arrangements.

Appendix II: CHILDES Transcript Coding Conventions

CHILDES 1.1 December 18, 1985

These codes are intended as an implementation of the CHILDES coding initiative agreed upon by the CHILDES Advisory Board in 1984 and sketched out in the MacWhinney and Snow article in the Journal of Child Language in 1985. The current codes were written by Brian MacWhinney and Kathy Marengo with assistance and comments from Kim Plunkett, Catherine Snow, Paul Fletcher, Jon Miller, Robin Chapman, Leonid Spektor, and Antonio Zampolli.

Changes from CHILDES 1.0 to CHILDES 1.1

1. Elimination of tier coding within the text line. For those who find that tier coding within the line is necessary, we have written a program called maketier.c which takes tier codings out of the text line.
2. Changed Head/ and Epi/ to simply @, so that Head/Age becomes @Age.
3. Replaced the distinction between headers and epiheaders with a distinction between constant and changeable headers.
4. Sharpened rules on punctuation of final quotations.
5. Eliminated the need for coding of $RFS and $FP on the error/hesitation tier.
6. Eliminated the use of commas.
7. Clarified the fact that $ISR (initial segment repetition) does not require the presence of the [/] symbol.
8. Noted that the xxx symbol can be placed anywhere on the text tier. When xxx is alone or at the end of an utterance it must be followed by a delimiter.
9. Added the ! symbol to identify tertiary tiers.

The coding conventions of the CHILDES system specify the following aspects of a transcript:

1. AUXILIARY FILES--files that document the corpus as a whole, such as 00readme.doc, or 00codes.doc.
2. HEADERS--information that describes facts that characterize the file as a whole or major sections of the file.
3. MAIN TEXTUAL LINE--the basic orthographic/morphemic representation of the interaction.
4. DEPENDENT TIERS--additional ways of coding the main line.
5. SPECIAL SYMBOLS--these are the symbols used to do the coding in (4) and (5) above.
6. SCOPING CONVENTIONS--ways of relating the tiers of (5) to the basic line of (4).

Auxiliary Files

Each corpus or collection of files should be accompanied by one or more documentation files.

1. 00readme.doc In this file there should be general information on the project, the subjects, the goals of the study, foci of coding, weaknesses in coding and transcription, special codes, and the like. This file should include references to cite when acknowledging use of the data and any special restrictions on use.

2. 00words.doc In this file there should be a list of special lexical forms used in the transcripts. These special forms are either rare standard words or words of the special lexicons discussed below such as motherese, familial, onomatopoeic forms and the like. They are indicated with the @m, @f, @o and other endings discussed later. This file should have data on each word in this order:

   orthography of special form
   part of speech of special form
   meaning of special form
   phonology of special form

3. 00codes.doc Any project-specific codes.

Headers

Headers have the following form:
@Attribute: Entry

Headers can be either constant or changeable. Constant headers characterize information that is necessarily constant throughout a given file. For example, the name of a child or the child's sex cannot change in the middle of a file. On the other hand, information about the situation or about tape location can change repeatedly in a file.

Constant Headers:

@Begin: This header should be placed at the very beginning of the file in order to guarantee that nothing has been lost during file copying or reformatting.

@End: This is the only header that should be placed at the end of the file. It should be placed as the very last line of the file to guarantee that nothing has been lost.

@Filename: Name of the file.

@Coding: Name and version of the coding scheme being used. The current scheme is CHILDES 1.1

@Participants: XXX Name Role, XXX Name Role, ... , XXX Name Role Here XXX represents the first three letters of the participants first name or some other three-letter code for the participant. Name is the participant's name. Role is his role. Whereever possible use these standard roles:

Teacher	Mother	Father	Target Child
Sister	Brother	Playmate	Grandmother
Grandfather	Babysitter	Investigator	

@Age of XXX: Age in years, months, and days for the three-letter code given in the head/participant record. For example, if one child is Bill and he is two years, three months and 12 days old, we have @Age of BIL: 2-3-12. Typically, this will only be coded for the target child, but in principle it can be given for any participant.

@SES of XXX: Standard qualifiers would be:

Lower	Working	Lower-Middle	Welfare
Middle	Upper-Middle	Upper	

@Education of XXX:
Grade in school from 0 to 20, where the numbers after 12 indicate years of college.

@Sex of XXX: Male or Female

@Birth of XXX: Dates in headers and elsewhere should be in the following form: 25-AUG-85, i.e day-month-year, using capitalized three-letter English abbreviations for the months: JAN, FEB, MAR, APR, MAY, JUN, JUL, AUG, SEP, OCT, NOV, DEC.

Changeable Headers:

@Date: As above

@Timing: If there is time coding relative to some event, the status of the marker event should be given. The usual marker event would be the beginning of the tape. If timing is absolute by the time of the day, that should be stated here.

@Situation: Use some of these standard situations whereever possible:

Breakfast	Lunch	Dinner	Bath
Washing-dishes	Returning-home		
Book-reading	Game-playing	Singing	Party
Holiday	Religious	School	Working
Learning-at-home	Outing (Camping, Hiking, etc.)		
Playmates-come-over	Visiting-playmates		
Cleaning-room	Getting-ready-to-go-out		

@Activities: This would provide further detail on the component activities in the situations.

@Location: (i.e. city, state etc.)

@Room Layout:

@Transcriber:

@Tape Location: (specify Tape ID, side, and footage)

If a new header attribute must be created, simply use the above format and create a new attribute code word.

Main Textual Line

The main textual line is given in conventional orthographic representation. However, the first word of utterances is not capitalized. Otherwise, the standard rules apply. This means that only proper nouns and the word "I" and its contractions are capitalized. Words beginning sentences are in lower case. Periods should be eliminated from abbreviations, so that "E.T." becomes "ET" and "Mrs." becomes "Mrs" without the final period.

The main textual line may be supplemented by certain characters that allow it to serve also as a morphemic representation. These are the symbols #, -, and 0 discussed below. The main textual line should be thought of as a string of lexical items. This means that phonological irregularities should not be represented on the main line, but should be coded on a phonetic coding tier or the error and hestitation tier. For example, stuttering or initial segment repetition as in "te-te-telephone" should be represented as simply "telephone" on the text line and coded as having initial segment repetition on the error-hesitation tier line.

Each textual "line" must begin with an asterisk. Then comes the first three letters of the speaker's name in all capital letters followed by a colon, i.e. *JIM: After the colon insert two spaces (this allows for use of the tab character) and begin the text. Text may be continued onto subsequent lines without using the speaker's three-letter name again, but those lines should be indented one by one 8-space tab. Example:

*JIM: We wish that all of you could have come with us to the store, but we just didn't have the time.

Dependent Tiers

Many additional tiers can be coded in relation to the main textual line. It is important to emphasize that we do not expect any researcher to code all files across all of these tiers. Rather, it is our intention to provide a common set of coding conventions which will allow the investigator to correctly represent those aspects of the data that are most important.

The $ symbol is used with qualifiers or specific codes within the tiers. For some tiers we do not plan to have a series of codes. These tiers simply use natural language descriptions. This type of coding will be called "standard language description coding." All tiers (other than text tiers and tertiary tiers) should begin with % and should be in lower case letters. If there is more than one entry within a single tier heading then they should be placed on the same line and separated by commas. The possible set of tiers is open-ended. However, the tiers most frequently used include these:

%pho: phonetic transcription. See the chapters in this issue on the proposed MULTIBET and UNIBET conventions for ASCII translations of the extended IPA system and for the standard phonemes of English. At the end of the MULTIBET proposal we also discuss the use of tertiary tiers dependent on the phonetic tier for coding suprasegmental information. Such tertiary tiers are marked with an exclamation mark at the beginning, as in !glo: for the tier coding glottal action.

%spa: speech acts (examples: echo, imitation)

*JIM: want some banana.
%spa: $IMIT

The speech act coding systems of Wells and Gutfreund and Ninio and Wheeler appear in Transcript Analysis (Vol 2, No 2 and Vol 3, No 1). Catherine Snow and others are working on further development of the coding for this tier.

%sit: situational coding. Standard language description coding.

%tim: time stamp coding. Time is given in hours-minutes-seconds, as in 12-20-55. Milliseconds can be coded after a decimal points, as in 12-20-55.567. Time can be either absolute (time of the day) or relative to any other event, which can be specified in a header.

%alt: alternative words.

*JIM: I need some cookies.
%alt: <2-4> can't see mommy

This line may also be used to provide a best guess for material represented as xxx or yyy.

%mor: morphemic semantics. The coding for this tier is based on Christian Lehmann's "Directions for Interlinear Morphemic Translations" from Folia Linguistica. The details of the application of this system to child language must be worked out language by language. When coding the morphemic tier, it is necessary to morphemicize the main textual tier, using the symbols # and - discussed below in "special symbols." In order to allow this tier to provide both morphemicization and translation for languages other than English, stems should be given in a slash notation with the English translation followed by the native language lemma (i.e. allomorph for searches). Thus, the Spanish word "vamos" [= we go] should be coded as:
%mor: go/ir#3PP

%par: paralinguistic features. (laugh, whisper, cry, yell, whimper, sneezes, speaks softly etc.) examples:

*JIM: mommy.
%par: whimper

*MOT: come on.
%par: laughs

%int: intonation.

*JIM: I want some milk now.
%int: <3> lengthening <4-5> fall

Codes for this tier have not yet been developed. However, consult Svartvik et al. "Survey of

Spoken English" for some conventions.

%exp: explanation.

*JIM: that that.
%exp: that's a red chair

Explanations can also be given on the text line by using the [=] symbol, as in: *JIM: that that. [=chair]

The deictic identity of objects can be specified in this way too.

%pos: part of speech. The parts-of-speech codes used by the Brown University and Lancaster University tagging projects are given in Transcript Analysis Vol 2 No 2 and are available on-line. Additional codes can be added as needed for special cases and other vocabulary. However, in order to maximize compatibility with the Brown codes, additional codes should be placed after the closest corresponding Brown code. For example, the word "everyone" would be coded as a nominal pronoun (NP) in the Brown system. In order to also code it as a "scope determiner" (SD), one should use this code: $NP/$SD

%phs: phrase structure notation - bracket words into phrases such as $NP or $VP. No standard set of phrase structure codes has yet been determined. However, the codes of LARSP are a good initial candidate set.

%err: errors and hesitations. We hope to be able to code errors and hesitations on a single tier. Joseph Stemberger (University of Minnesota) is currently working on the design of these codes.

initial phoneme repetition - $IPR
*PET: I said that.
%err: <2> $IPR
multiple word repetition - $MWR
word repetition - $WR
initial syllable repetition - $ISR
N.B. all repetitions require use of the [/] symbol in the text line, as discussed below
drawling - $D

%eng: gloss using English morphemes corresponding to morphemes in transcript

%act: actions - people doing something. Standard language description coding.

%gpx: gestures - proxemic. Standard language description coding.

%sta: staging directions, who talks to whom. Standard language description coding.

%com: comments (multiple comments string out on one comment line)

%cod: the catch-all tier for general coding or anything that does not fit in the above. This can also be used when you want to mix codes into a single tier for economy or ease of entry. However, it may then be necessary later to separate these codes into separate tiers.

Special Symbols

The first five symbols are used to delimit utterances. Commas are not used on the text line.

. the end of an unmarked (declarative) utterance

! the end of an exclamation

? the end of a question

[...] trailing off of an utterance

[\] interruption by the other speaker

[+] completion of the utterance of the previous speaker, as in

*HEL: If Bill had known [\]
*WIN: [+] he would have come.

[#] an unfilled pause. The length of the pause can be given following the # in seconds.milliseconds, as in [#5.44] which is a pause lasting 5.44 seconds or [#333] which is a pause lasting 333 seconds (5 minutes and 33 seconds). To code a long pause without estimating the time, use [#long].

www placed on the text tier for material which is not necessarily meaningless, but which is not transcribed for technical reasons. This could occur, for example, when the speaker produces material in a language the transcriber doesn't know or doesn't know how to transcribe.

*TUN: www.
%eng: the sun

Note that the symbols www, xxx, yyy, and zzz should be followed by utterance delimiters when they end utterances.

xxx unintelligible speech (use lower case xxx)

yyy speech for which transcriber provided a phonetic reading but no orthographic reading. This symbol should be used when the transcriber cannot relate the sounds to any lexical items.

zzz paralinguistic material or vocal qualifiers such as screams, crying, etc. which cannot be transcribed in any standard fashion.

" marks the beginning and the end of a direct quotation

[?] previous word unclear

[!] previous word stressed

[-] utterance broken for tier coding. This is placed both at the end of the line broken off and at the beginning of the continuation line.

[//] self interruption with both retracing and correction, as in: *BET: I wanted [//] I thought I wanted to invite Margie.
Do not code $RFS (retraced false start) on the error line, since this would be misleadingly redundant.

[/] self interruption with retracing without correction, as in:

*BET: I wanted [/] I wanted to invite Margie.

' contraction

[=text] following text gloss or explains meaning of previous word. When a longer explanation is necessary, use the %exp tier line.

0 placed on text tier when there is no utterance but another tier line is necessary

i.e. *MOT: 0.
%act: goes to the front door

< >[overlap follows] the material in angle brackets is the discourse of the first speaker overlapped by the second speaker

< >[overlap precedes]
the material in angle brackets is the discourse of the second speaker overlapped by the first speaker

on the text tier line, prefixes and suffixes are preceded by a number sign. i.e. dog#s

- morphemically segmented text lines should have stems preceded by dashes when there are prefixes (e.g. #un-decipher#able)

or to analyse compounds (e.g. storm-trooper)

or to represent rote forms (e.g. put-this ball)

0 omission of a morpheme

that go here that go#0es[3S] here
he here he 0is here

@ Non-standard lexical items are indicated by attaching the @ symbol plus a further letter to the end of a morpheme, for example:

@b	babbling	abame@b	"babbling"
@c	child-invented form	gumma@c	"sticky cereal"
@f	familial form	mumsa@f	"mother"
@p	phonological consistant form	aga@p	"pointing morph"
@m	motherese (baby-talk form)	owie@m	"small hurt"
@s	second-language form	istenem@s	"my God"
@o	onomatopoeic	bowwow@o	"dogs barking"
@l	letter name	w@l	"the letter w"
@q	qualifier	uhnuh@q	"no"
@h	hesitation form	uh@h	"uh"
@e	extended English	squoze@e	"squeezed"

Scoping or Coreferencing Conventions

1. referring to whole utterance, no markers needed, if only one utterance is coded after the * sign.
2. referring to a whole utterance, when several utterances are coded after an * sign, use <2u> for the second utterance.
3. referring to the whole turn as a point
   a. <bef> what follows was done before the utterance
      *JIM: ooooh.
      %act: <bef> runs down the hall
   b. <dur> what follows was done during the utterance
      *JIM: what's this.
      %ges: <dur> points to shoes
   c. <aft> what follows was done after the utterance
      *MOT: go upstairs please.
      *JIM: ok
      %com: <aft> break in interaction
      [this shows that the comment occured after Jim had spoken and does not refer to his utterance]
4. referring to a stretch of words use <2-4> for the second, third, and fourth words. Do not count special symbols like #, but do count xxx as one word.
5. referring to a word as a point, indicate point in one of these ways:
   a. <.3> = before the third word or
   b. <2.> = after the second word or
   c. <3> = during the third word

   These scopings can be combined in this way:
   a. <5,10> = on the fifth and tenth words
   b. <.4,5.>= before the fourth and after the fifth words
   c. <.3-8> = from before the third to through the eighth word

   referring to a segment as a point, first identify the word and then the segment, using the letter "s", as in:
   a. <5s3> = during the third segment of the fifth word
   b. <5s3-6>= from the third to the sixth segment of the fifth word
   c. <5s.4,8s4> = before the fourth segment of the fifth word and during the fourth segment of the eighth word

6. The above may be combined with utterance identification in this way:
   a. <2u3-3u4> = from the third word of the second utterance to the fourth word of the third utterance
   b. <2u-6u> = from the second to the sixth utterance
   c. <2u5s4> = second utterance fifth word fourth segment Note that this is could be useful for speech act codings within turns, since several utterances may serve a given function.
   d. <2u3-5> = from the third to fifth word of the second utterance